

本周尝试对交易数据使用 Seasonal-Trend Decomposition 进行分解，并验证结果的有效性。提取 10 月 8 日凌晨 0 点的 45 万条交易数据。

1.初步分析

首先对前 15 分钟的数据的每秒交易量，交易金额，交易平均值进行简单的统计。

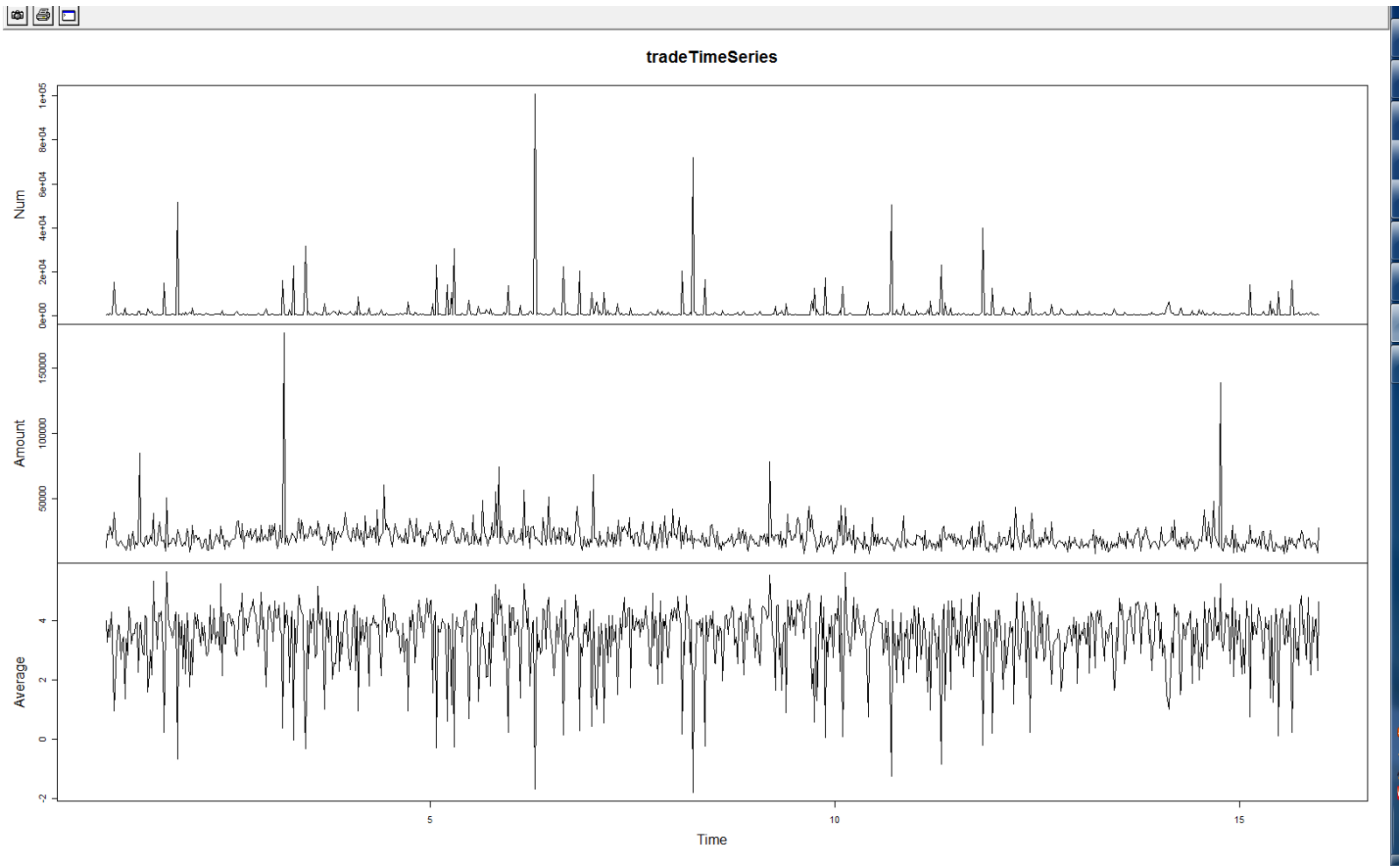


图 1 前 15 分钟的交易数据。从上到下表示交易数目，交易金额，交易平均价格的变化。对交易平均价格进行了 STL 分解（Seasonal-Trend Decomposition）。采用 R 语言实现。

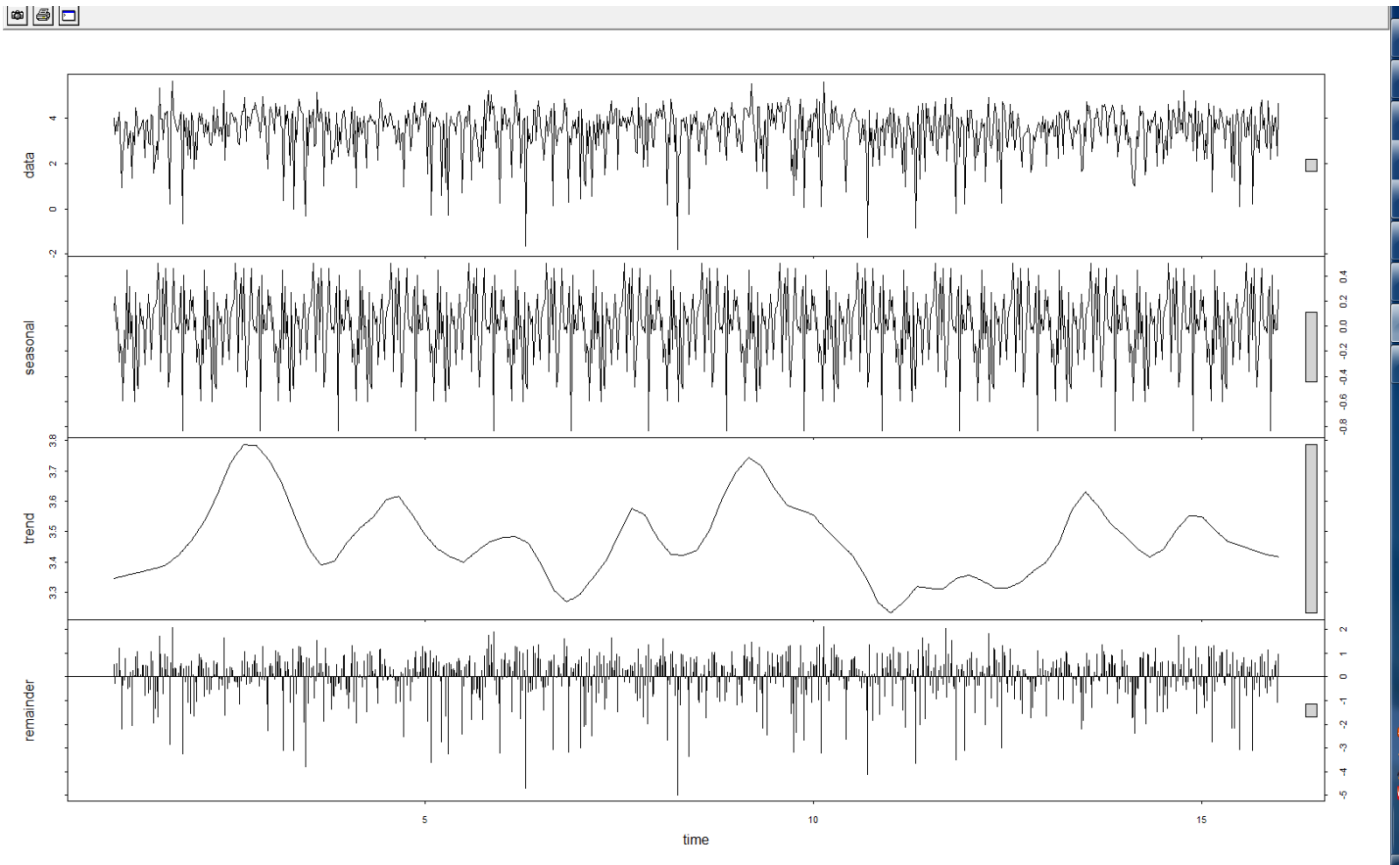


图 2 对 15 分钟内的取对数后的交易平均价格的 STL 分解。

从上到下四幅图表示：原始数据，Seasonal 项，Trend 项，Reminder 项。

从 SEASONAL 的形态可以得出一个结论：交易数据并不具有严格的周期特征。

从 TREND 中可以发现一些信息，比如：在 4 分钟，7 分钟，11 分钟左右出现了一波均值较低的交易。

从 Reminder 可以直接发现某一时刻出现的异常交易。但是残余比较没有规律，而且与原始数据十分相似。

使用平均交易金额直接进行 STL 分解效果不是很好而且不是很实用。因此需要结合数据的其他维度，比如用户 ID，交易类目来分析出异常交易的信息。

普通的 STL 处理的是单值的时序数据，我们则需要对高维的交易信息进行处理，数据的维度也是一个挑战。

2.高维数据的简化

每一笔交易 T 可以看成是一个 13 个元素组成的向量：

$T = \{ \text{ali_date}, \text{buyer_id}, \text{seller_id}, \text{auction_price}, \text{buy_amount}, \text{aa_city}, \text{aa_prov}, \text{cat_id}, \text{cat_name}, \text{cat1}, \text{name1}, \text{lgo_prov}, \text{lgo_city} \}$;

依次代表：交易时间，卖家 ID，买家 ID，单笔交易价格，单笔交易数量，卖家城市，卖家省份，叶子类目 ID，叶子类目名称，大类目 ID，大类目名称，买家城市，买家省份。

假设卖/买家的城市在一天内是不变的，因此从卖/买家 ID 就知道其城市，从而也可以知道其省份。交易的类目 ID 对应于其中文名称。如果先对大类目进行分析，可以省略其叶子类目。因此，一笔交易 T 就可以简化为六个维度组成的向量：

$T = \{ \text{ali_date}, \text{buyer_id}, \text{seller_id}, \text{auction_price}, \text{buy_amount}, \text{cat1} \}$

3.交易概率模型

上周组会报告的文章《Spatiotemporal Social Media Analytics for Abnormal Event Detection and Examination using Seasonal-Trend Decomposition》中，作者使用 LDA 对文本数据进行处理得到文本的主题，再使用 STL 对一段时间内某个主题出现的次数进行分析。对于交易数据，问题在于交易比较杂乱，没有类似于文本的特定“主题”以供抽取。

LDA 的方法在处理此处的高维交易数据时可能不是很有效。因此我尝试利用概率模型来分析每分钟的异常交易情况，首先作出以下两个假设：

假设 1： 对于一个给定的类目 C，其商品的单价 V 服从正态分布。

即 $V \sim N(\mu, \sigma)$ ，其概率函数为： $F(V; \mu, \sigma)$ 。

则属于类目 C 的单价为 V 的交易是正常交易的置信水平为： $1 - 2 \times |F(V; \mu, \sigma) - 0.5|$ 。

之所以做出这个假设，是因为同一类商品的价值可以认为大致相似。而不同商家的价格基于商品价值，根据经济学的价值规律：“价格围绕价值上下波动正是价值规律作用的表现形式。”比如手机的价格大概都在千元左右浮动。因此就采用了比较直观的高斯分布。

假设 2： 对于某一天的交易，单位时间内某一特定卖家或买家的 ID 进行交易的次数 k 服从参数为 λ 泊松分布。

即 $P \sim \pi(\lambda)$ ，其概率为 $P(X=k)$ 。

商品交易概率分布的参数估计：

在前面两个假设的基础上，对商品的价格分布，以及用户的购买次数的分布进行参数估计。

1).特定类目的价格分布

对于每个类目 C_m ($1 \leq m \leq 97$)，类目 C_m 的商品价格服从正态分布 $N_m(\mu_m, \sigma_m)$ 。首先统计 C_m 的商品的价格，使用极大似然估计估计出其分布参数 μ_m 与 σ_m 。

$$\mu^* = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

公式 1

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

公式 2

此处 n 表示类目 C_m 的交易总数， x_i 表示第 i 笔类目 C_m 交易的交易金额。

2).用户交易次数的分布

统计出全部 ID 的交易频率。单位时间内某 ID 交易次数服从泊松分布 $\pi(\lambda)$ ，同样使用极大似然估计估计出其参数 λ :

$$\hat{\lambda}_{MLE} = \frac{1}{n} \sum_{i=1}^n k_i.$$

公式 3

其中 n 表示交易的 ID 的数目， k_i 表示第 i 个用户在单位时间内的交易次数。

4.异常值定义与分析:

交易的异常程度直接与交易的频率与交易的金额相关。因此，对于某一笔交易 T_i ($1 \leq i \leq 4.5 \times 10^5$)，其买家或卖家 ID 在一段时间内出现的交易次数为 k ，该笔交易的价格为 A ，该笔交易的商品数量为 n ，其交易类目为第 m 类， F_m 为该类目的商品价格的概率函数。 P 为泊松分布概率函数。其交易为正常交易的概率 p 为:

$$p(T_i) = P(X=k) \times (1 - 2 \times |F_m(A/n; \mu_m, \sigma_m) - 0.5|)$$

公式 4

以上公式简单的说，就是一段时间内 ID 出现 k 次的概率，与交易单价为 A/n 的置信水平的乘积。用以表示该交易是正常程度。

某一时刻的异常程度与该时刻内所有交易的异常程度有关。设该时刻内一共有 n 笔交易。因此，定义某一时刻 t 异常值 $P_{abnormal}(t)$ 为:

$$P_{abnormal}(t) = \sum_i^n (1 - p(T_i))$$

公式 5

简单的说，就是用该时刻所有交易的异常概率的和，来表示该时刻总体交易的异常程度。

以秒为单位，计算得到 15 分钟内每个时刻的异常值，如下图:

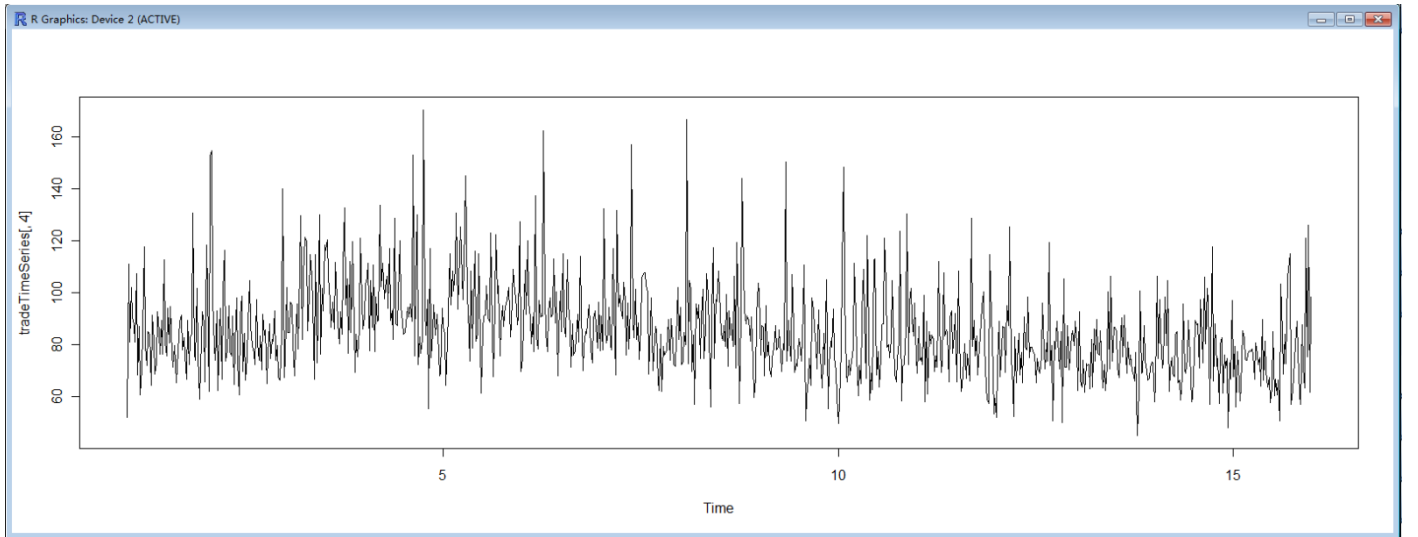


图 3 按照上述方法计算的 10 月 8 日 0 点前 15 分钟的异常值 $P_{abnormal}$ 曲线。

对 $P_{abnormal}$ 进行 STL 分解可得:

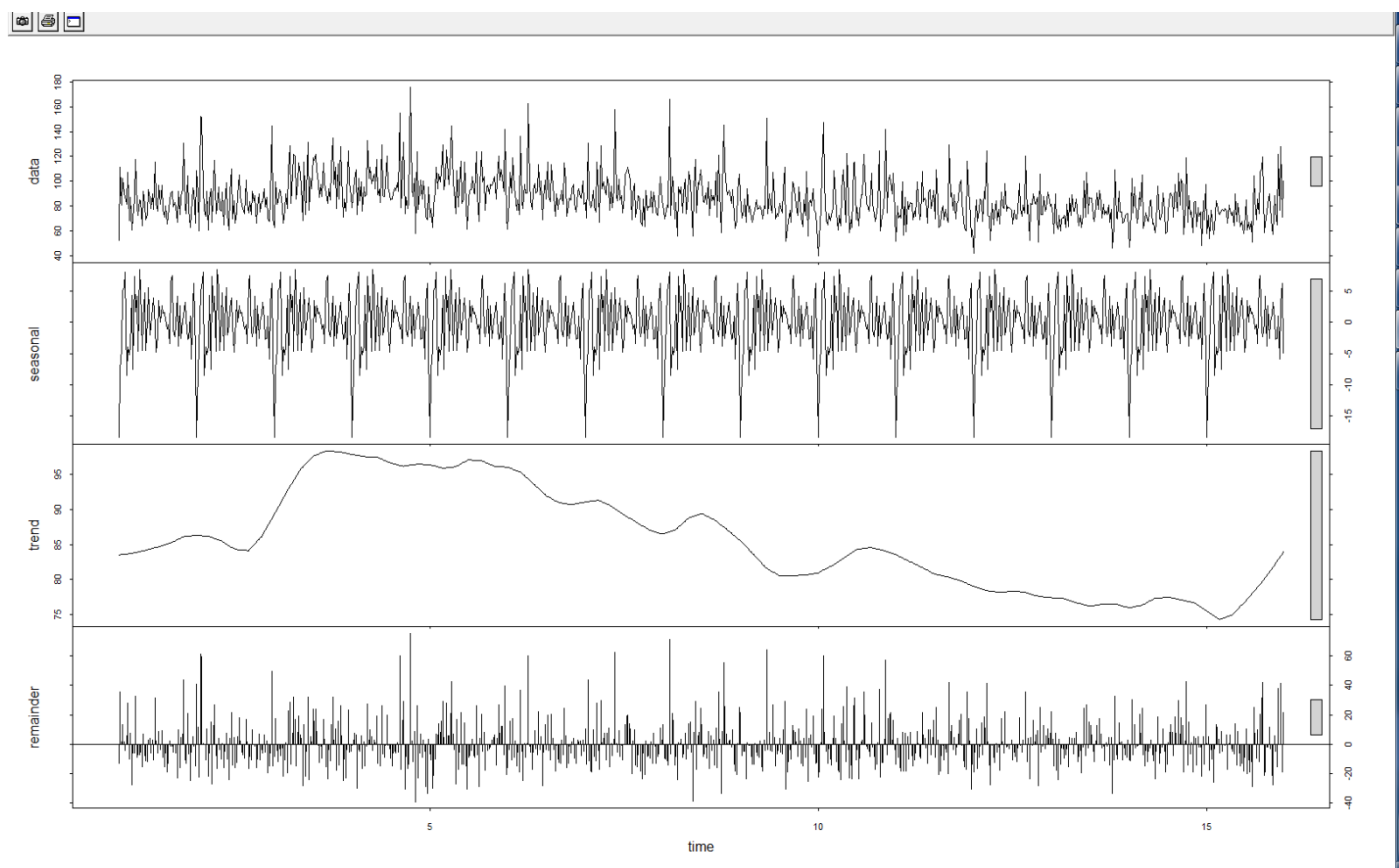


图 4 对 15 分钟内的异常值的 STL 分解。从上到下四幅图表示：原始数据，Seasonal 项，Trend 项，Reminder 项。

如图 4 可以看出其 seasonal 与 reminder 两项仍然是比较杂乱没有什么规律。因此，基本可以认为在普通情况下，一个小时以内的淘宝交易本身就是不具周期规律的一种数据。

值得注意的是 trend 项提供了交易异常值的趋势信息。可以看出其值在 3 分钟到 7 分钟处于一个较高水平。可以进一步对 3-7 分钟的数据进行的分析。

5.数据验证与猜想：

为了寻求对上述方法的验证，在原始数据中查找异常交易。仅仅通过眼睛在海量数据中寻找还是有一点麻烦，因此将 ID 进行统计，很容易发现了以下异常交易：

A328441		0:03:34											
	A	B	C	D	E	F	G	H	I	J	K	L	M
328418	0:03:32	806795856	737695153	8.94	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328419	0:03:32	912361577	737695153	8	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328420	0:03:32	1016069718	737695153	9.6	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328421	0:03:33	303680050	737695153	4.99	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328422	0:03:33	108773801	737695153	8.95	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328423	0:03:33	805646506	737695153	4.76	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328424	0:03:33	479997499	737695153	4.99	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328425	0:03:33	1047091734	737695153	5.99	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328426	0:03:33	885022615	737695153	3.99	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328427	0:03:33	808056006	737695153	3.2	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328428	0:03:33	838025812	737695153	4.99	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328429	0:03:33	36737309	737695153	9.9	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328430	0:03:33	806795856	737695153	3.2	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328431	0:03:33	808056006	737695153	9.6	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328432	0:03:33	803152549	737695153	5.6	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328433	0:03:33	912361577	737695153	8	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328434	0:03:33	840124594	737695153	5.6	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328435	0:03:33	71054248	737695153	4.2	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328436	0:03:34	48552548	737695153	3.99	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328437	0:03:34	885022615	737695153	7.2	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328438	0:03:34	88335073	737695153	1.99	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328439	0:03:34	287384261	737695153	5.25	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328440	0:03:34	1041632939	737695153	5.99	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328441	0:03:34	179929056	737695153	5.25	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328442	0:03:34	95545408	737695153	1.99	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328443	0:03:35	1053747011	737695153	3.99	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328444	0:03:35	836355814	737695153	3.99	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328445	0:03:35	801742709	737695153	4.99	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328446	0:03:35	69479823	737695153	8.66	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328447	0:03:35	86361621	737695153	5	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328448	0:03:35	1047312201	737695153	4.99	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328449	0:03:35	1041632939	737695153	3.9	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328450	0:03:35	807167481	737695153	3.99	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328451	0:03:35	1024248841	737695153	8	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328452	0:03:35	807167481	737695153	4.99	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328453	0:03:36	1024248841	737695153	4.99	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328454	0:03:36	60721148	737695153	3.99	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328455	0:03:36	909424865	737695153	16	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328456	0:03:36	1047804955	737695153	4.99	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328457	0:03:36	57604200	737695153	1.99	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328458	0:03:36	801742709	737695153	7.5	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328459	0:03:36	1047312201	737695153	3.99	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328460	0:03:36	807972356	737695153	8.94	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328461	0:03:36	380758247	737695153	8	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N
328462	0:03:36	411642123	737695153	3.99	1	杭州	浙江		33	书籍/杂志	33	书籍/杂志	\N

图 5 买家 ID 为“737695153”在一段时间内进行了多笔异常交易。A 列表示交易的时间；B,C 两列表示卖价买家 ID；D, E 两列表示交易价格与交易的数量；F,G 两列表示卖家交易地点信息；H-K 列表示交易的类目的信息最后两列表示买家交易地点信息。

很容易发现可疑的买家 ID“737695153”。

此买家的交易行为与我们之前所预计的不一样的地方是：原本以为交易金额会比较小，该买家所有的交易金额处于正常水平，并未出现 1 分钱以下的交易。

买家地点未知“\N”。更为奇怪的是虽然对应的其卖家 ID 各不相同，但是其地点都是位于浙江杭州。

这样的交易模式与我们先前猜想的 0 元交易，单一卖、买家的交易还是有些不同的。

异常数据可能的解释：

- 猜想 1：某个专职刷信誉的用户的交易行为。他为一个杭州的卖书团体（依据：卖家 ID 不同，地点相同）刷信誉。
 猜想 2：可能是商品质量检查部门的交易行为。他们伪装成买家以普通渠道购买商品（依据：购买类目相同，交易金额无异常），抽查杭州的电子书商的卖书质量。